Identifying Anatomical Concepts in Biomedical Text for Automatic Selection of Images

Powell J. Bernhardt[a], Thomas C. Rindflesch[b], Halil Kilicoglu[b], Michele Tringali[c]

[a]*Temple University, Philadelphia, PA, USA*
[b]*National Library of Medicine , Bethesda, MD, USA*
[c]*Azienda Ospedaliera Santa Maria della Misericordia, Udine, Italy*

Abstract

Images can provide a valuable complement to online information resources. We suggest a method for automatically suggesting appropriate illustrations for anatomically-oriented text. The system we are developing uses natural language processing and domain knowledge in the Unified Medical Language System® (UMLS), ® relying particularly on 'part of' information in an anatomical ontology. We discuss the application of our methodology to a set of gastrointestinal endoscopy reports and provide examples linking text to images in several online image repositories.

Keywords:

Natural language processing, anatomical images, Unified Medical Language System

Introduction

We recently began research to connect the language of biomedical text with images illustrating the semantic content of that text. Although images provide a valuable complement to text-based information sources, significant challenges are involved in affording effective access to image databases.

In information management research, attention is paid to various aspects of accommodating images, including indexing through direct computation on images [1], [2], [3] and the development of methods for representing image content for effective retrieval [4], [5]. A variety of methods address ways that the retrieval of images differs from retrieval of text, regardless of the indexing methodology [6], [7], [8] [9]. Navigational aids are being developed [10] for image databases, and recently, the interpretation of captions associated with images in the biomedical research literature has been addressed [11].

Our research complements this work by focusing on automatically selecting illustrative images for anatomical text based on the UMLS knowledge sources [12] and natural language processing to characterize the semantic content of text. In the initial phase of this project we concentrate on the automatic illustration of gastrointestinal endoscopy (GIE) reports with appropriate anatomical images. We are developing our methods by processing a set of (anonymized) GIE reports from Clarkson University Hospital, Omaha, Nebraska.

Anatomy is often viewed as foundational to the life sciences, and digital anatomical image datasets, such as the Visible Human Project® [13], are being developed to provide access to anatomical resources for health care professionals and researchers as well as students and the public. Our research facilitates this access by seeking to provide a seamless connection between online text and images.

There are several possible approaches to automatically determining appropriate anatomical illustrations for biomedical text. One method might simply provide an image for each anatomical concept found in the text. However, this is not an optimal solution when several parts of a coherent structure are discussed. Nor is it ideal when the parts of a structure are mentioned but the structure itself is not explicitly articulated. In both situations it would be better to have an illustration of the entire structure.

These phenomena prevail in GIE reports. For example, anatomical concepts are highlighted in the text in (1), taken from the description of findings section of a report.

1. The endoscope was introduced past the **cricopharyngeus** down through the **esophagus** and into the **stomach**. **Pylorus** was identified and intubated. **Second portion of the duodenum** and **duodenal bulb** were examined.

Rather than providing a separate image for each anatomical concept in (1), illustrations reflecting the fact that the cricopharyngeus is part of the pharynx (not mentioned) and that the pylorus is part of the stomach would be more informative. Further, when the entire text is considered, an image of the gut (with labeled parts) would be most useful.

In this paper we discuss a method for suggesting optimal illustrations for clinical text based on natural language processing that provides support for determining how the anatomical concepts mentioned in a section of text interact in a 'part of' hierarchy. We draw on natural language processing tools being developed at the National Library of Medicine as well as the medical domain knowledge contained in the UMLS Metathesaurus.® Of particular importance for this project is the University of Washington Digital Anatomist Symbolic Knowledge Base (UWDA) [14], [15], one of the constituent sources in the Metathesaurus. This is an anatomical ontology that includes entities arranged in taxonomic ('isa') as well as meronymic ('part of') hierarchies.

Methods

Processing to connect images to text falls conveniently into two phases: a) text-to-ontology processing and b) ontology-to-image processing. The first phase identifies anatomical concepts in text and determines which concepts should be used to select the best image to illustrate the content of that text. Ontology-to-image processing is concerned with locating images for the concepts identified in the first step; this involves the characteristics of the images contained in a particular database and how they are indexed. Although our primary focus in this paper is text-to-ontology processing, we include some discussion of the issues involved in accessing online anatomical images.

Text-to-Ontology Processing

In this project, text-to-ontology processing is concerned with determining the anatomical content of GIE reports, which relies on two notions: a) categorization of the concepts involved and b) the 'part of' relations that obtain among the structures named by those concepts.

Each Metathesaurus concept is categorized by one or more semantic types, such as 'Phamacologic Substance' and 'Disease or Syndrome'. Semantic types have been merged into coarser grained semantic groups [16], and the semantic group for anatomy includes 'Anatomical Structure', 'Body Location or Region', 'Body Part, Organ, or Organ Component', 'Body Space or Junction', 'Body Substance', 'Body System', 'Cell', 'Cell Component', 'Embryonic Structure', 'Fully Formed Anatomical Structure', and 'Tissue'.

Concepts occurring in the UWDA ontology are situated in a 'part of' hierarchy as appropriate. For example, stomach is 'part of' the foregut which is 'part of' the gut.

We first map input text to the Metathesaurus and then limit concepts to those that have a semantic type in the Anatomy group and that occur in UWDA. Such concepts are submitted to further processing in the gastroenterology domain, and the 'part of' hierarchy is then used to group them into coherent structures that form the basis for suggesting images.

Selecting anatomical concepts

Our natural language processing to support image selection is based on previous work applied to anatomically-oriented text [17], [18]. Input is subjected to look-up in the SPECIALIST Lexicon [19] and a stochastic tagger [20] to resolve part-of-speech ambiguity. An underspecified syntactic parser then provides simple noun phrases to MetaMap [21], which matches text to concepts in the Metathesaurus.

Further processing is applied to anatomical concepts in order to resolve ambiguity, context-specific meaning, and missing synonyms in the gastroenterology domain [22]. For example, *rugae* is mapped to the concept "Gastric rugae" and *antrum* to "Pyloric antrum." The term *mid esophagus* (which does not occur in the Metathesaurus) is resolved as "Middle third of esophagus."

Anatomical concepts identified by this processing are highlighted in (2).

2. DESCRIPTION OF FINDINGS: The patient was placed in the left lateral decubitus position. The upper endoscope was inserted into the **esophagus** and advanced to 32 cm where the patient's **anastomosis** was noted. This appeared widely patent. The endoscope was advanced into the **stomach** which showed part of the **stomach** above the diaphragmatic hiatus. The endoscope was easily advanced through the **pylorus**, into the **duodenum**.

Non-anatomical Metathesaurus concepts, such as "Left lateral decubitus position," with semantic type 'Spatial Concept', and "Endoscope ('Medical Device') were identified by MetaMap but eliminated. Although diaphragmatic hiatus is an anatomical notion, this term does not occur in the Metathesaurus and should be mapped to the concept "Hiatus esophageus of diaphragm" in this domain.

Determining 'part of' relations

Processing aimed at selecting appropriate anatomical images begins by examining the anatomical concepts identified in the previous step. As suggested above, the distribution of these concepts in the UWDA 'part of' hierarchy is an important component in selecting an appropriate anatomical image for that text. This distribution is computed by first determining the line of ascent in the 'part of' hierarchy for each relevant concept. A line of ascent includes the concept of interest as well as all direct ancestors. For example, the lines of ascent for the anatomical concepts in (2) are shown in Figure 1. Although "Anastomosis" is in the Anatomy group, it does not appear in the UWDA 'part of' hierarchy.

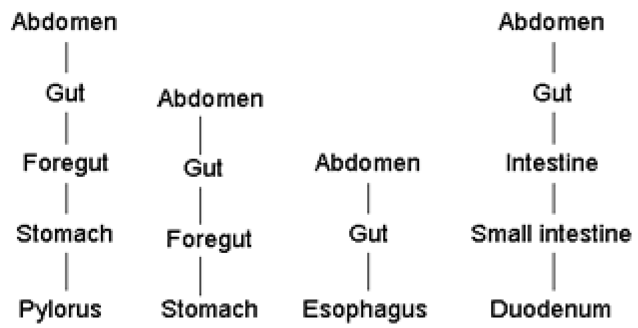

Figure 1. Lines of ascent for anatomical concepts in (2)

Once lines of ascent have been determined, they are grouped into "families" sharing a common ancestor, and the lowest common ancestor is computed for the family with the largest number of members. For example, the line of ascent for "Esophagus" is [Esophagus, Gut, Abdomen] while the line of ascent for "Stomach" is [Stomach, Foregut, Gut, Abdomen].

The lowest common ancestor for these concepts is “Gut.” The lowest common ancestor and its descendants are shown in tree form in Figure 2 for the largest family from the lines of ascent in Figure 1.

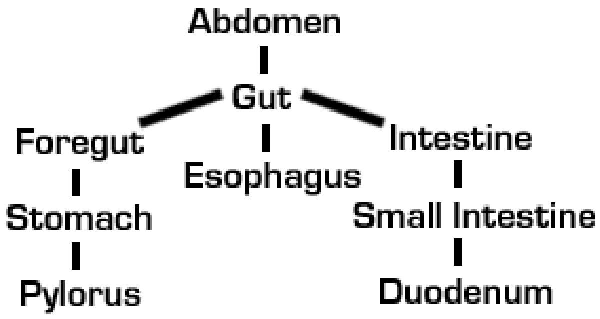

Figure 2. Lowest common ‘part of’ ancestor (gut) for the concepts in (2)

An image that represents a coherent section of a GIE report should be broad enough to incorporate all relevant terminology, and we suggest that an image of the lowest common ancestor of the largest family satisfies this requirement. Appropriate illustration should also display anatomical detail. For this, images of the descendants of the lowest common ancestor can be displayed. In the current example, the user can be given the option of viewing the larger structure (e.g. gut) or detailed images of a specific location or region of that structure (e.g. stomach or esophagus).

Ontology-to-Image Processing

After the lowest common ancestor and its descendants have been identified, it would be ideal to search the Internet and return images that correspond to those concepts; however, effective retrieval depends on the characteristics of online image databases. Such databases contain images of diverse content based on various points of view of anatomy. Some contain macroscopic images of varying level of detail either in the form of drawings or photographs of cadavers, while others contain microscopic or endoscopic images from various body systems. Further, anatomical image databases may be indexed using various features, including terms from a specific controlled vocabulary. In order to access these resources effectively it is necessary to know these characteristics ahead of time.

Choosing the most useful anatomical images as illustrations for text depends on the needs of the user in addition to the semantic content of the text. Gross anatomical images with labeled structures may be appropriate for students and consumers. However, a specialist may only be interested in gastrointestinal endoscopic images, for example. As a pilot project, we are exploring methods for providing images from several anatomical points of view to illustrate the content of a document.

The Visible Human database is indexed with UWDA concepts and has an associated mechanism for finding and retrieving images, AnatQuest [23]. However, currently, available images are limited to the thorax.

We manually selected three online repositories containing gastrointestinal images. The Medical Encyclopedia in MEDLINE*plus* contains drawings of gross anatomy images from A.D.A.M., Inc. GASTROLAB, a collection of selected gastrointestinal endoscopic images, is indexed with terms from the Metathesaurus Version of Minimal Standard Terminology Digestive Endoscopy (MTHMST) terminology. Finally, Bristol Biomedical Image Archive has micrographs indexed with Medical Subject Headings (MeSH).

In order to facilitate an automatic link between the output from our semantic content processing and these three repositories of images, we created an ancillary database for each online resource. The ancillary databases have two fields, one for an indexing term and another for a URL accessing an image of that concept. In each local database the indexing terms are drawn from the vocabulary appropriate to the online image repository. For example, the entries in our database for the endoscopy images are shown in (3), where the entries in the first field are terms from the MTHMST vocabulary.

3. Pharynx|http://www.gastrolab.net/ya469x2.jpg
   Pylorus|http://www.gastrolab.net/ya048h.jpg
   Duodenum|http://www.gastrolab.net/ya406x.jpg

Preliminary Evaluation

Given the exploratory nature of this research attempting to automatically suggest the best images for anatomically-oriented text, we have so far not conducted a formal evaluation of the software being developed. Informal evaluation suggests that although identification of anatomical concepts is generally accurate, some problems remain. For example, the system currently does not recognize any anatomical concepts in the sentence, *Examination reveals normal bulb, second, and third portion*, which refers to the following concepts: “Superior part of duodenum,” “Descending part of duodenum,” and “Horizontal part of duodenum.” In order to identify these concepts the term *bulb* should be resolved to the “Superior part of duodenum.” Further, and more challenging, processing is required to determine that the terms *second* and *third portion* refer to the appropriate parts of the duodenum.

We have not yet addressed the more difficult issue of evaluating the appropriateness of the images suggested by the system as illustrations for a given text. Such evaluation depends on several variables in addition to the accuracy of the identification of anatomical concepts. These include characteristics of the available images as well as the user’s level of interest and point of view.

An Example

There are several steps that must be completed before images are accessible to a user. Once the lines of ascent and lowest common ancestor have been found, we use the UMLS Knowledge Source Server [24] to find corresponding MeSH and MTHMST terms for each concept, which allows access to the indexing terms in our ancillary databases. The user can then choose a concept to be illustrated and specify a point of view, macro-, micro-, or endoscopic. The program then looks up the up the corresponding entry in the appropriate ancillary database and resolves the URL to retrieve an image. For example, semantic content output for the input text from (2) above is given in (4).

4. [Duodenum,Small intestine,Intestine,Gut,Abdomen]
   [Esophagus,Foregut,Gut,Abdomen]
   [Pylorus,Stomach,Foregut,Gut,Abdomen]
   [Stomach,Foregut,Gut,Abdomen]

   Lowest Common Ancestor: Gut

Example illustrations available for these concepts are given below. Figure 3 is a macroscopic image of the gut from A.D.A.M., Inc. from the MEDLINE*plus* Medical Encyclopedia; Figure 4 is an endoscopic view of the pylorus from GASTROLAB; and Figure 5 is a micrograph of the duodenum from the Bristol Biomedical Image Archive.

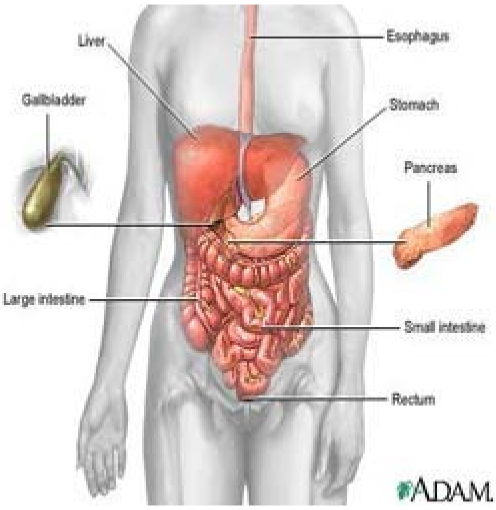

Figure 3. Macroscopic image of the gut

Figure 4. Endoscopic image of the pylorus

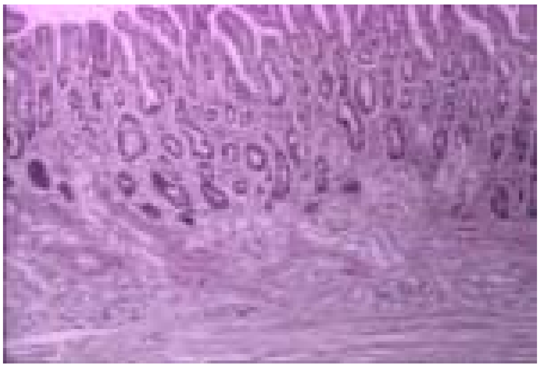

Figure 5. Microscopic image of the duodenum

Conclusion

We have discussed a methodology for automatically suggesting appropriate images to illustrate anatomically-oriented text. Concentrating on gastrointestinal endoscopy reports, we use natural language processing to map text to concepts in the UMLS. In particular, the UWDA 'part of' hierarchy is used to guide the selection of the best image for a text. We provide examples linking concepts to images in several online image repositories.

Ackowledgments

We are grateful to Dr. Timothy McCashland of the University of Nebraska Medical Center for making the GIE reports available to us.

References

[1] Fung CY, Kia FL. A new approach for image classification and retrieval. In: Hearst M, Gey F, and Tong R , eds. Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, 1999; p. 301.

[2] Mattie ME, Staib L, Stratmann E, Tagare H, Duncan J, Miller PL. PathMaster: Content-based cell image retrieval using automated feature extraction. Journal of the American Medical Informatics Association 2000: 7(4):404-15.

[3] Antani S, Long LR, Thoma G, Stanley RJ. Vertebra Shape Classification using MLP for Content-Based Image Retrieval. Proc. IJCNN International Joint Conference on Neural Networks, 2003; pp. 160-65.

[4] Bidgood WD, Bray B, Brown N, Rossi Mori A, Spackman KA, et al. Image acquisition context: Procedure description attributes for clinically relevant indexing and selective retrieval of biomedical images. Journal of the American Medical Informatics Association 1999: 6(1):61-75.

[5] Quintana Y. Organization and retrieval in a pictorial digital library. In: Allen RB and Rasmussen E, eds. Proceedings of the 2nd ACM International Conference on Digital Libraries, 1997; pp. 13-20.

[6] Aslandogan Y, Their C, Yu CT, Zou J, Rishe N. Using semantic contents and WordNet in image retrieval. In: Belkin NJ, Narasimhalu AD, and Willett P, eds. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1997; pp. 286-95.

[7] Harmandas V, Sanderson M, Dunlop MD. Image retrieval by hypertext links. In: Belkin NJ, Narasimhalu AD, and Willett P, eds. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1997; pp. 296-303.

[8] Ravela S, Manmatha R. Image retrieval by appearance. In: Belkin NJ, Narasimhalu AD, and Willett P, eds. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1997; pp. 278-85.

[9] Tagare HD, Jaffe CC, Duncan J. Medical image databases: A content-based retrieval approach. Journal of the American Medical Informatics Association 1997: 4(3):184-98.

[10] Wacholder N, Venuti JM, Krauthammer M, and Molholt P. Designing a navigational ontology for browsing and accessing 3D anatomical images. In Overhage JM, ed. Proceedings of the AMIA Annual Symposium, 2000; p. 1152.

[11] Cohen WW, Wang R, Murphy RF. Understanding Captions in Biomedical Publications. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2003; pp. 499-504.

[12] Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical language System: An informatics research collaboration. Journal of the American Medical Informatics Association 1998: 5(1):1-13.

[13] Ackerman MJ. Accessing the Visible Human Project. D-Lib Magazine, October 1995 (http://www.dlib.org/dlib/october95/10ackerman.html).

[14] Brinkley JF, Rosse C. The Digital Anatomist distributed framework and its application to knowledge-based medical imaging. Journal of the American Medical Informatics Association 1997: 4(3):184-98.

[15] Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, and Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: The Digital Anatomist Symbolic Knowledge Base. Journal of the American Medical Informatics Association 1998: 5(1):17-40.

[16] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001: 10(Pt 1):216-20.

[17] Sneiderman CA, Rindflesch TC, and Bean CA. Identification of anatomical terminology in medical text. Proceedings of the AMIA Annual Symposium, 1998; pp. 428-32.

[18] Rindflesch TC, Bean CA, Sneiderman CA. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proceedings of the AMIA Annual Symposium, 2000; pp. 704-8.

[19] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proceedings of the Annual Symposium on Computing Applications in Medical Care, 1994; pp. 235-9.

[20] Cutting D, Kupiec J, Pedersen J and Sibun P. A practical part-of-speech tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, 1992.

[21] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proceedings of the AMIA Annual Symposium, 2001; pp. 17-21.

[22] Tringali M, Rindflesch TC, Kilicoglu H, Fiszman M, Bodenreider B 2003. Strategies for mapping gastrointestinal endoscopy reports to UMLS Metathesaurus concepts. Submitted to MEDINFO.

[23] Anatomical images for the public: A report to the Board of Scientific Counselors. Lister Hill National Center for Biomedical Communications, National Library of Medicine. May, 2003.

[24] Bangalore A, Thorn KE, Tilley C, Peters L. The UMLS Knowledge Source Server: An object model for delivering UMLS data. Proceedings of the AMIA Annual Symposium, 2003.

Address for correspondence

Powell J. Bernhardt

231 Hedgerow Drive

Swedesboro, NJ, USA